Statistics and Computing

W.N. Venables
B.D. Ripley

Modern Applied Statistics with S-Plus

With 124 Figures

Springer Science+Business Media, LLC

W.N. Venables
Department of Statistics
University of Adelaide
Adelaide, South Australia 5005
Australia

B.D. Ripley
Professor of Applied Statistics
University of Oxford
1 South Parks Road
Oxford OX1 3TG
England

Series Editors:

W. Eddy
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
USA

W. Härdle
Institut für Statistik und Ökonometrie
Humboldt-Universität zu Berlin
Spandauer Str. 1
D-10178 Berlin
Germany

S. Sheaffer
Australian Graduate School of Management
PO Box 1
Kensington
New South Wales 2033
Australia

L. Tierney
School of Statistics
University of Minnesota
Vincent Hall
Minneapolis, MN 55455
USA

Library of Congress Cataloging-in-Publication Data
Modern applied statistics with S-PLUS / W.N. Venables, B.D. Ripley.
p. cm. -- (Statistics and computing)
Includes bibliographical references and index.
ISBN 978-1-4899-2821-4 ISBN 978-1-4899-2819-1 (eBook)
DOI 10.1007/978-1-4899-2819-1
1. S-Plus. 2. Statistics--Data processing. 3.Mathematical statistics--Data processing. I. Ripley, Brian D., 1952-
II. Title. III. Series.
QA276.4.V46 1994
005.369--dc20 94-21589

Printed on acid-free paper.

Production managed by Jim Harbison; manufacturing supervised by Jacqui Ashri.
Photocomposed pages prepared from the authors' PostScript files.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4899-2821-4

Preface

Increases in computer power and falling costs have changed dramatically the concept of the 'statistician's calculator'. From 1903 to 1963 this would have been mechanical or electro-mechanical supplemented by a slide rule. In 1973 the HP-35 scientific calculator was available to the wealthy. By 1983 we used BASIC on a microcomputer or a statistical package on a time-sharing central service, but graphical output was cumbersome if available at all. By 1993 a statistician could have a workstation displaying high-resolution graphics with megaflops of computing power and hundreds of megabytes of disc storage. In this era the statistical system S has flourished.

It is perhaps misleading to call S a statistical system. It was developed at AT&T's Bell Laboratories as a flexible environment for data analysis, and in its raw form implements rather few statistical concepts. It does however provide the tools to implement rather elegantly many statistical ideas, and because it can be extended by user-written procedures (both in its own language and in C and FORTRAN), it has been used as the basis of a number of statistics systems, notably S-PLUS from the StatSci division of MathSoft, which has implemented a number of topics (such as time series and survival analysis) using the S language. (The S system is now marketed exclusively by StatSci.)

The user community has contributed a wealth of software within S. This book shows its reader how to extend S, and uses this facility to discuss procedures not implemented in S-PLUS, thereby providing fairly extensive examples of how this might be done.

This is not a text in statistical theory, but does cover modern statistical methodology. Each chapter summarizes the methods discussed, in order to set out the notation and the precise method implemented in S. (It will help if the reader has a basic knowledge of the topic of the chapter, but several chapters have been successfully used for specialized courses in statistical methods.) Our aim is rather to show how we analyse datasets using our 'statistical calculator' S-PLUS. In doing so we aim both to show how S can be used and how the availability of a powerful and graphical system has altered the way we approach data analysis and allows penetrating analyses to be performed routinely. Once calculation became easy, the statistician's energies could be devoted to understanding his or her dataset.

The core S language is not very large, but it is quite different from most other statistics systems. We describe the language in some detail in the early chapters, but these are probably best skimmed at first reading; Chapter 1 contains the most basic ideas, and each of Chapters 2 and 3 are divided into 'basic' and 'advanced'

sections. Once the philosophy of the language is grasped, its consistency and logical design will be appreciated.

The chapters on applying S to statistical problems are largely self-contained, although Chapter 6 describes the language used for linear models, and this is used in several later chapters. We expect that most readers will want to pick and choose amongst the later chapters, although they do provide a number of examples of S programming, and so may be of interest to S programmers not interested in the statistical material of the chapter.

This book is intended both for would-be users of S-PLUS (or S) as an introductory guide and for class use. The level of course for which it is suitable differs from country to country, but would generally range from the upper years of an undergraduate course (especially the early chapters) to Masters' level. (For example, almost all the material is covered in the M.Sc. in Applied Statistics at Oxford.) Exercises are provided only for Chapters 2–5, since otherwise these would detract from the best exercise of all, using S to study datasets with which the reader is familiar. Our librariy provides many datasets, some of which are not used in the text but are there to provide source material for exercises. (We would stress that it is preferable to use examples from the user's own subject area and experience where this is possible.)

Both authors take responsibility for the whole book, but Bill Venables was the lead author for Chapters 1–4, 6, 7 and 9, and Brian Ripley for Chapters 5, 8 and 10–15. (The ordering of authors reflects the ordering of these chapters.)

The datasets and software used in this book are available for electronic distribution from sites in the USA, England and Australia. Details of how to obtain the software are given in Appendix A. They remain the property of the authors or of the original source, but may be freely distributed provided the source is acknowledged. We have tested the software as widely as we are able (including under both Unix and Windows versions of S-PLUS), but it is inevitable that system dependencies will arise. We are unlikely to be in a position to assist with such problems.

The authors may be contacted by electronic mail as

```
venables@stats.adelaide.edu.au
ripley@stats.ox.ac.uk
```

and would appreciate being informed of errors and improvements to the contents of this book.

Acknowledgements:

This book would not be possible without the S environment which has been principally developed by Rick Becker, John Chambers and Allan Wilks, with substantial input from Doug Bates, Bill Cleveland, Trevor Hastie and Daryl Pregibon. The code for survival analysis is the work of Terry Therneau. The S-PLUS code is the work of a much larger team acknowledged in the manuals for that system.

We are grateful to the many people who have read and commented on one or more chapters and who have helped us test the software, as well as to those whose problems have contributed to our understanding and indirectly to examples and exercises. We can not name all, but in particular we would like to thank Adrian Bowman, Michael Carstensen, Sue Clancy, David Cox, Anthony Davison, Peter Diggle, Matthew Eagle, Nils Hjort, Richard Huggens, Francis Marriott, David Smith, Patty Solomon and StatSci for the provision of a very early copy of S-PLUS 3.2 for final testing of our material.

Bill Venables
Brian Ripley
April 1994

Contents

Preface v

Typographical Conventions xiii

1 Introduction 1

1.1 A quick overview of S 3
1.2 Getting started 4
1.3 Bailing out 6
1.4 Getting help with functions and features 7
1.5 An introductory session 8
1.6 What next? 16

2 The S Language 17

2.1 A concise description of S objects 17
2.2 Calling conventions for functions 25
2.3 Arithmetical expressions 26
2.4 Reading data 32
2.5 Finding S objects 36
2.6 Character vector operations 38
2.7 Indexing vectors, matrices and arrays 40
2.8 Matrix operations 45
2.9 Functions operating on factors and lists 51
2.10 Input/Output facilities 54
2.11 Customizing your S environment 57
2.12 History and audit trails 59
2.13 Exercises 59

3 Graphical Output 61

3.1 Graphics devices 61
3.2 Basic plotting functions 65
3.3 Enhancing plots 70

3.4 Conditioning plots 74
3.5 Fine control of graphics 76
3.6 Exercises 83

4 Programming in S **85**
4.1 Control structures 85
4.2 Writing your own functions 90
4.3 Finding errors 97
4.4 Calling the operating system 103
4.5 Some more advanced features. Recursion and frames 105
4.6 Generic functions and object-oriented programming 110
4.7 Using C and FORTRAN routines 113
4.8 Exercises 119

5 Distributions and Data Summaries **121**
5.1 Probability distributions 121
5.2 Generating random data 123
5.3 Data summaries 125
5.4 Classical univariate statistics 129
5.5 Density estimation 134
5.6 Bootstrap and permutation methods 141
5.7 Exercises 146

6 Linear Statistical Models **147**
6.1 A linear regression example 147
6.2 Model formulae 153
6.3 Regression diagnostics 157
6.4 Safe prediction 161
6.5 Factorial designs and designed experiments 162
6.6 An unbalanced four-way layout 169
6.7 Multistratum models 177

7 Generalized Linear Models **183**
7.1 Functions for generalized linear modelling 187
7.2 Binomial data 189
7.3 Poisson models 196
7.4 A negative binomial family 200

8 Robust Statistics **203**
8.1 Univariate samples 204
8.2 Median polish 210
8.3 Robust regression 212
8.4 Resistant regression 217
8.5 Multivariate location and scale 222

9 Non-linear Regression Models **223**
9.1 Fitting non-linear regression models 224
9.2 Parametrized data frames 226
9.3 Using function derivative information 226
9.4 Non-linear fitted model objects and method functions 229
9.5 Taking advantage of linear parameters 230
9.6 Examples 231
9.7 Assessing the linear approximation: profiles 237
9.8 General minimization and maximum likelihood estimation . . . 239

10 Modern Regression **247**
10.1 Additive models and scatterplot smoothers 247
10.2 Projection-pursuit regression 255
10.3 Response transformation models 258
10.4 Neural networks 261
10.5 Conclusions 265

11 Survival Analysis **267**
11.1 Estimators of survivor curves 269
11.2 Parametric models 273
11.3 Cox proportional hazards model 279
11.4 Further examples 285
11.5 Expected survival rates 298
11.6 Superseded functions 299

12 Multivariate Analysis **301**
12.1 Graphical methods 301
12.2 Cluster analysis 311
12.3 Discriminant analysis 315
12.4 An example: *Leptograpsus variegatus* crabs 322

13 Tree-based Methods **329**
13.1 Partitioning methods 330
13.2 Cutting trees down to size 342
13.3 Low birth weights revisited 345

14 Time Series **349**
14.1 Second-order summaries 352
14.2 ARIMA models 361
14.3 Seasonality 367
14.4 Multiple time series 373
14.5 Nottingham temperature data 376
14.6 Other time-series functions 380
14.7 Backwards compatibility 382

15 Spatial Statistics **383**
15.1 Interpolation and kriging 383
15.2 Point process analysis 392

References **397**

Appendices

A Datasets and Software **407**
A.1 Directories 408
A.2 Sources of machine-readable versions 411
A.3 Caveat 412

B Common S-PLUS Functions **413**

C S versus S-PLUS **429**

D Using S Libraries **431**
D.1 Creating a library 433
D.2 Sources of libraries 434

E Command Line Editing **437**

F Answers to Selected Exercises **439**

Index **447**

Typographical Conventions

Throughout this book S language constructs and commands to the operating system are set in a monospaced typewriter font `like this`.

We often use the prompts `$` for the operating system (it is the standard prompt for the Unix Bourne shell) and `>` for S-PLUS. However, we do *not* use prompts for continuation lines, which are indicated by indentation. One reason for this is that the length of line available to use in a book column is less than that of a standard terminal window, so we have had to break lines which were not broken at the terminal.

Some of the S-PLUS output has been edited. Where complete lines are omitted, these are usually indicated by

```
....
```

in listings; however most *blank* lines have been silently removed. Much of the S-PLUS output was generated with the options settings

```
options(width=65, digits=5)
```

in effect, whereas the defaults are `80` or `87` (depending on the release) and `7`. Not all S-PLUS functions consult these settings, so on occasion we have had to manually reduce the precision to more sensible values.